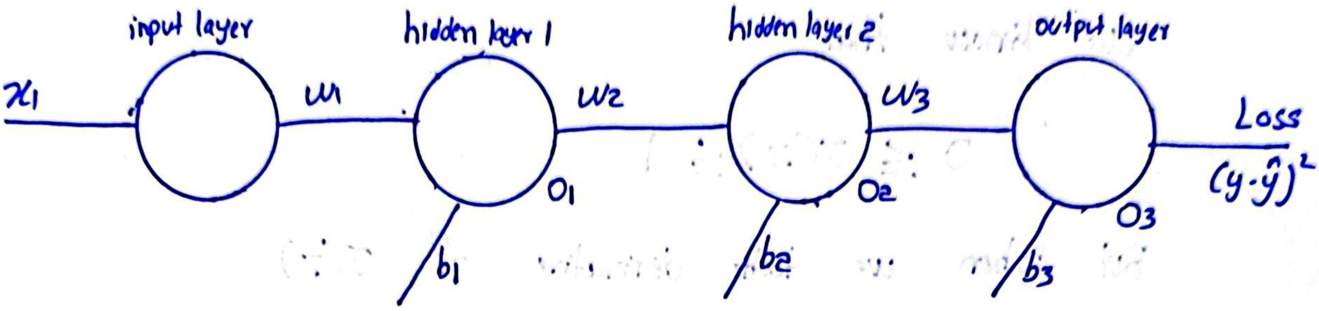
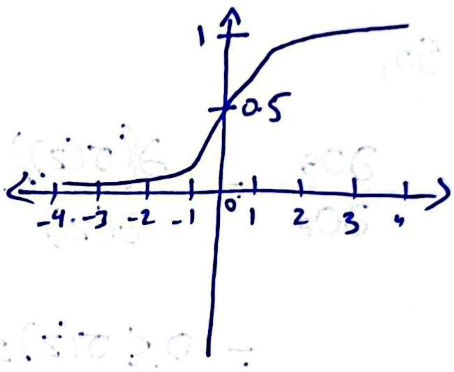
Vanishing Gradient Problem



Let's use sigmoid activation function

$$z = \sum_{i=1}^{n} w_i x_i + b$$

$$\sigma(z) \Rightarrow 0 \text{ to } 1$$



$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$w_{1_{new}} = w_{1_{old}} - \alpha \frac{\partial L}{\partial w_{1_{old}}}$$

α 0.01

$$\frac{\partial L}{\partial w_{1_{old}}} = \frac{\partial L}{\partial O_3} \times \frac{\partial O_3}{\partial O_2} \times \frac{\partial O_2}{\partial O_1} \times \frac{\partial O_1}{\partial w_{1_{old}}}$$

$$O_3 = \sigma(w_3 \times O_2 + b_3)$$

Let $z = w_3 \times O_2 + b_3$

$$O_3 = \sigma(z)$$

$$\frac{\partial O_3}{\partial O_2} = \frac{\partial(\sigma(z))}{\partial(z)} \times \frac{\partial(z)}{\partial O_2}$$

We know that

$$0 \leq \sigma(z) \leq 1$$

But when we take derivative of $\sigma(z)$

$$0 \leq \sigma(z) \leq 0.25$$

So,

$$\frac{\partial O_3}{\partial O_2} = \frac{\partial(\sigma(z))}{\partial(z)} \times \frac{\partial(z)}{\partial O_2}$$

$$= 0 \leq \sigma(z) \leq 0.25 \times \frac{\partial(w_3 \times O_2 + b_3)}{\partial O_2}$$

$$\frac{\partial O_3}{\partial O_2} = 0 \leq \sigma(z) \leq 0.25 \times w_{3_{old}}$$

This will effect

$$w_{1_{new}} = w_{1_{old}} - \alpha \boxed{\frac{\partial L}{\partial w_{1_{old}}}} \Rightarrow \text{small value}$$

$$w_{1_{new}} \approx w_{1_{old}}$$

This will cause vanishing gradient problem

Vanishing Gradient

It occurs when the gradients used to update the neural network become very small as they propagated back through the network. This can lead to minimal or no updates to the weights in initial layer.